MIRACLE at GeoCLEF Query Parsing 2007: Extraction and Classification of Geographical Information

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3],
José Carlos González-Cristóbal[1,3], and José Miguel Goñi-Menoyo[1]

[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.
slana@diatel.upm.es, jvillena@it.uc3m.es,
josecarlos.gonzalez@upm.es, josemiguel.goni@upm.es

Abstract. This paper describes the participation of MIRACLE research consortium at the Query Parsing task of GeoCLEF 2007. Our system is composed of three main modules. The first one is the Named Geo-entity Identifier, whose objective is to perform the geo-entity identification and tagging, i.e., to extract the "where" component of the geographical query, if there is any. Then, the Query Analyzer parses this tagged query to identify the "what" and "geo-relation" components by means of a rule-based grammar. Finally, a two-level multiclassifier first decides whether the query is indeed a geographical query and, should it be positive, then determines the query type according to the type of information that the user is supposed to be looking for: map, yellow page or information.

Keywords: Linguistic Engineering, classification, geographical IR, geographical entity recognition, gazetteer, Geonames, tagging, query classifier, WordNet.

1 Introduction

MIRACLE team is a research consortium formed by research groups of three different Spanish universities (Universidad Politécnica de Madrid, Universidad Autónoma de Madrid and Universidad Carlos III de Madrid) along with DAEDALUS, a private company founded as a spin-off of these groups and a leading company in the field of linguistic technologies in Spain. MIRACLE has taken part in CLEF since 2003 in most tracks and tasks, including the main bilingual, monolingual and cross lingual tasks [1] as well as in ImageCLEF, WebCLEF, GeoCLEF [2] [3] and Question Answering tracks.

This paper describes the MIRACLE participation [4] at the Query Parsing task of GeoCLEF 2007 [5]. In the following sections, we will first give an overview of the architecture of our system. Afterwards we will elaborate on the different modules. Finally, the results will be presented and analyzed.

C. Peters et al. (Eds.): CLEF 2007, LNCS 5152, pp. 786–793, 2008.

2 System Description

The system architecture is shown in Figure 1. Note that our approach consists of three sequential tasks executed by independent modules [4]:

- **Named Geo-entity Identifier:** performs the geo-entity identification and a query expansion with geographical information.
- **Query Analyzer:** identifies the "what" and "geo-relation" components of a geographical query.
- **Query Type Classifier:** determines the type of geographical query.

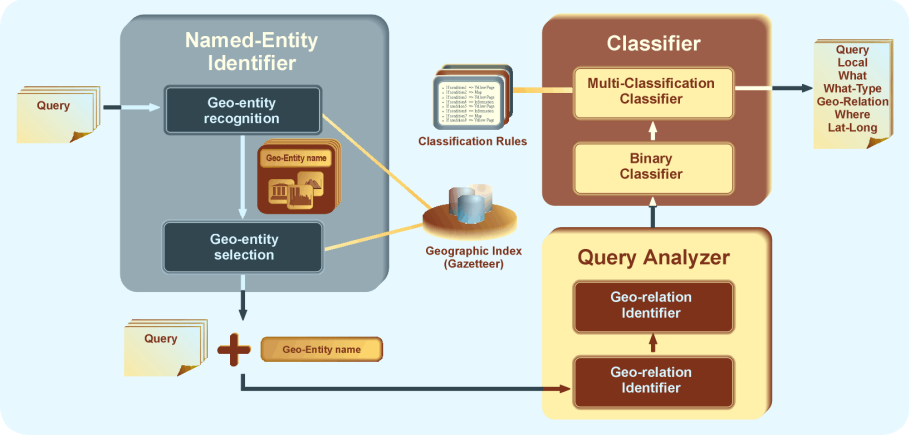

Fig. 1. Overview of the system

2.1 Named Geo-entity Identifier

The objective of this module is to perform the geo-entity identification and tagging, i.e., to extract the "where" component of the query, should there be any. It is composed of two main components: a gazetteer, i.e. a database with geographical resources that constitutes the knowledge base of the system, and a geo-entity parser built on top of it.

Our gazetteer is built up from the Geonames geographical database [6], available free of charge for download under a Creative Commons attribution license. It contains over 8 million geographical names with more than 6.5 million unique features about 2.2 million populated places and 1.8 million alternate names. Those features include a unique identifier, the resource name, alternative names (in other languages), county/region, administrative divisions, country, continent, longitude, latitude, population, elevation and timezone. All features are categorized into one out of 9 feature classes and further subcategorized into one out of 645 feature codes. Geonames integrates geographical data (such as names of places in various languages, elevation or population) from various sources, mainly the Geonet Names Server (GNS) [7] gazetteer of the National Geospatial Intelligence Agency (NGA), the Geographic Names Information System (GNIS) [8] gazetteer of the U.S. Geographic Survey, the GTOPO30 [9]

digital elevation model for the world developed by United States Geological Survey (USGS) and Wikipedia, among others.

For our purposes, all data was loaded and indexed in a MySQL database, although not all fields (such as time zone or elevation) are used: the relevant fields are UFI (unique identifier), NAME_ASCII (name), NAME_ALTERNATE (alternate names), COUNTRY, ADM1 and ADM2 (administrative region where the entity is located), FEATURE_CLASS, FEATURE_TYPE, POPULATION, LATITUDE and LONGITUDE. To simplify the query processing, each row is complemented with the expansion of country codes (ES→Spain) and/or state codes (NC→North Carolina) – when applicable. The final database uses 865KB.

The geo-entity parser carries out the following tasks:

- **Geo-entity recognition:** identifies named geo-entities [3] using the information stored in the gazetteer, looking for candidate named entities matching any sub-string of one or more words [10] included in the query and not included in a stop-word (or stop-phrase) list [11].
  The stopword list is mainly automatically built by extracting those words that are both common nouns and also georeference entities, assuming that the user is asking for the common noun sense (for example, "Aguilera" –for "Christina Aguilera"– or "tanga" – "thong"). Specifically we have used lexicons for English, Spanish, French, Italian, Portuguese and German, and have selected words that appear at least with a certain frequency in the query collection. The stopword list currently contains 1,712 entries.
- **Geo-entity selection:** The selected named geo-entity will be the one with the longest number of matching words and, if the same, the one with higher score. The score is computed according to the type of geographic resource (country, region, county, city…) and its population, as shown in the following table.

Table 1. Entity score

Feature type	Code	Score
Capital and other big cities	PPLA, PPLC, PPLG	Population+100,000,000
Political entities	PCL, PCLD, PCLF, PCLI, PCLIX, PCLS	Population+10,000,000
Countries	A	Population+1,000,000
Other cities	PP, STLMT	Population+100,000
Other	*	Population
For all cities, if country/state name/code is also in the query	PP, STLMT	Score += 100,000,000

  Those values where arbitrarily chosen after different manual executions and subsequent analysis.
- **Query tagging:** expands the query with information about the selected entity: name, country, longitude, latitude, and type of geographic resource.

The output of this module is the list of queries in which a possible named geo-entity has been detected, along with their complete tagging. Table 2 shows an example of a possible output.

Table 2. Example of tagged geo-entities

Query*| *score*| *ufi*| *entity*| *state (code)*| *country (code)*| *latitude*| *longitude*| *feature_class*| *feature_type
airport {{alicante}} car rental week| 2693959| 2521976| Alicante| | Spain (ES)| 38.5| -0.5| A| ADM2
bedroom apartments for sale in {{bulgaria}}| 10000000| 732800| | Bulgaria (BG)| 43.0| 25.0| A| PCLI
hotels in {{south lake tahoe}}| 123925| 5397664| South Lake Tahoe| California (CA)| United States (US)| 38.93| -119.98| P| PPL
Helicopter flight training in southwest {{florida}}| 100100000| 4920378| Florida|Indiana (IN)| United States (US)| 40.16| -85.71|P|PPL

Note that the geo-entity is specifically marked in the original query, enclosed between double curly brackets, to help the following module to identify the rest of the components of the geographical query.

2.2 Query Analyzer

This module parses each previously tagged query to identify the "what" and "geo-relation" components of a geographical query, sorting out the named geo-entity detected by the previous module, enclosed between curly brackets. It, in turn, consists of two subsystems:

- **Geo-relation identifier:** identifies and qualifies spatial relationships using rule-based regular expressions. Its output is the input list of queries expanded with information related to the identified "geo-relation".

  Table 3 shows the output of this module for the previous examples.

Table 3. Geo-relation expansion

Query*| *geo-relation*| *entity*| *state*| *country*| *country (code)*| *latitude*| *longitude*| *feature_class*| *feature_type
airport {{alicante}} car rental week| NONE| Alicante| | Spain| ES| 38.5|-0.5| A| ADM2
bedroom apartments for sale #@#IN#@# {{bulgaria}}| IN| Bulgaria| | Bulgaria| BG| 43.0| 25.0| A| PCLI
hotels #@#IN#@# {{south lake tahoe}}| IN| South Lake Tahoe| California| United States| US| 38.93| -119.98| P| PPL
helicopter flight training in #@#SOUTH_WEST_OF#@# {{florida}}| SOUTH_WEST_OF| Florida| Indiana| United States| US| 40.16| -85.71| P| PPL

Note that the geo-relation is also marked in the original query.

- **Concept identifier:** analyses the output of the previous step and extracts the "what" component of a geographical query applying manually defined grammar rules based on the identified "where" and "geo-relation" components.

2.3 Query Type Classifier

Finally, the last step is to decide whether the query is indeed a geographical query and, should it be positive, to determine the type of query, according to the type of information that the user is supposed to be looking for:

- **Map type:** users are looking for natural points of interest, such as rivers, beaches, mountains, monuments, etc.
- **Yellow page type:** businesses/organizations, like hotels, restaurants, hospitals, etc.
- **Information type:** users are looking for text information (news, articles, blogs).

The process is carried out by a two level classifier [4]:

1. **First level:** a binary classifier to determine whether a query is a geographical or a non-geographical query. This simple classifier is based on the assumption that a query is geographical if the "where" component is not empty.
2. **Second level:** a multi-classification rule-based classifier to determine the type of geographical query. The multi-classifier treats the tagged queries as a lexicon of semantically related terms (words, multi-words and query parts).
   The classification algorithm applies a knowledge base that consists on a set of manually defined grammar rules, including nouns and grammatically related part-of-speech categories as well as the type of geographical resource. The different valid lemmas are unified using Wordnet synsets [4].

3 Results

For the evaluation, multiple human editors labeled 500 queries that were chosen to represent the whole query set. Then all the submitted results were manually compared to those queries following a strict criterion where a match should have all fields correct. Table 4 shows the evaluation results of our submission, using the well-known evaluation measures of precision, recall and F1-score.

According to the task organizers [5], our submission achieved the best performance (F1-score) out of the 6 submissions of this year, which was satisfying, given our hard work. Other groups used similar approaches, but we think that the coverage of our

Table 4. Overall results

Precision(1)	Recall(2)	F1-score(3)
0.428	0.566	0.488

$$\text{precision} = \frac{\text{correctly\_tagged\_queries}}{\text{all\_tagged\_queries}} \tag{1}$$

$$\text{recall} = \frac{\text{correctly\_tagged\_queries}}{\text{all\_relevant\_queries}} \tag{2}$$

$$\text{F1} - \text{score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

gazetteer, an adequate stopword list, the algorithm for geo-entity selection and the precision of the query classifier let us make the difference with respect to other systems.

In addition, as participants in the task were provided with the evaluation data set, we have further evaluated our submission to separately study the results for each component of the geographical queries and also analyze the level-by-level performance of the final classifier.

Table 5 shows the individual analysis of the classifier per each field. The first-level classifier (LOCAL) achieves a precision of 75.40%, but the second-level classifier reduces this value to 56.20% for the WHAT-TYPE feature. According to a strict evaluation criterion, this would be the precision of the overall experiment.

Table 5. Individual analysis per field

	LOCAL		WHAT		WHAT-TYPE		WHERE		ALL	
	Total	%	Total	%	Total	%	Total	%	Total	%
All topics	377	75.40	323	64.60	281	56.20	321	64.20	259	51.80
Well-classified	377	100.00	323	85.67	281	74.53	321	85.15	259	68.70

However, if evaluated only over well-classified (geographical/non geographical) queries, the precision arises to 74.53% for the same feature. This great improvement shows that the precision of the system highly depends on the correct classification of the query and the first-level classifier turns out to be one of the key components of the system. The confusion matrix for this classifier is shown in Table 6, which shows that the precision is 73%. The conclusion for future participations is that more effort should be invested on improving this classifier to increase the overall performance.

Table 6. Confusion matrix for the binary classifier

	LOCAL YES	LOCAL NO
ASSIGNED YES	297	111
ASSIGNED NO	12	80

Precision(1)	Recall(2)	Accuracy(3)
0.73	0.96	0.75

$$(1)\ \text{precision} = \frac{TP}{TP + FP} \qquad (2)\ \text{recall} = \frac{TP}{TP + FN} \qquad (3)\ \text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 7 shows the same evaluation for the multiclassifier, but individualized per class and calculated over all topics. The lowest precision corresponds to "Yellow Page" queries. The explanation is that our gazetteer lacks that type of information such as names of hotels, hospitals, shopping centers, etc. This issue will be solved for future participations.

The following Table 8 shows the same evaluation per class, but calculated only over topics which are correctly classified by the first-level binary classifier. It is interesting to observe an increase in precision for all types of queries, but the relative distribution remains the same. As in the previous table, the lowest recall corresponds to "Map" queries. The difficulty to classify, parse and execute these queries may explain this fact.

Table 7. Evaluation of the multiclassifier, per class, all topics

Type	Precision	Recall	Accuracy
Yellow Page	0.43	0.95	0.61
Map	0.74	0.52	0.89
Information	0.93	0.20	0.88

Table 8. Evaluation of the multiclassifier, per class, correctly-classified topics

Type	Precision	Recall	Accuracy
Yellow Page	0.61	0.99	0.75
Map	0.92	0.55	0.89
Information	1.00	0.21	0.86

Last, we have to express some disagreements with the evaluation data provided by the organizers. Although some issues may be actual errors, most are due to the complexity and ambiguity of the queries. Table 9 shows some examples of queries that have been classified as geographical by our system but have been evaluated as false-positives. In fact, we think that it would be almost impossible to reach a complete agreement in the parsing or classification for every case among different human editors. The conclusion to be drawn from this is that the task to analyze and classify queries is very hard without a previous contact and without the possibility of interaction and feedback with the user.

Table 9. Some examples of ambiguities

QueryNo	Query	Extracted "where"	Why not?
113501	calabria chat	calabria, Italy	chat rooms about the region of Calabria?
443245	Machida	machida, Japan	Hiroko Machida (actress), Kumi Machida (artist) or the city of Machida?
486273	montserrat reporter	montserrat, Montserrat	online newspaper or reporters in Montserrat?

4 Conclusions and Future Work

According to a strict evaluation criterion where a match should have all fields correct, our system reaches a precision value of 42.8% and a recall of 56.6% and our submission is ranked 1st out of 6 participants in the task.

However, a detailed evaluation of the confusion matrixes reveals that some extra effort must be invested in "user-oriented" disambiguation techniques to improve the first level binary classifier for detecting geographical queries, as it is a key component to eliminate many false-positives.

In addition, the analysis of the confusion matrixes for the multiclassifier that are calculated over the topics correctly classified by the first level classifier shows that the probability that a geographical query is classified as "Yellow Page" is very high. This could be related to the uneven distribution of topics (almost 50% of the geographical queries belong to this class). In addition, "Information" type queries have a very low recall. These combined facts point out that the classification rules have not been able to establish a difference between both classes. We will focus on this issue in future participations.

Acknowledgements. This work has been partially supported by the Spanish R&D National Plan, by means of the project RIMMEL (Multilingual and Multimedia Information Retrieval, and its Evaluation), TIN2004-07588-C03-01; and by the Madrid's R&D Regional Plan, by means of the project MAVIR (Enhancing the Access and the Visibility of Networked Multilingual Information for the Community of Madrid), S-0505/TIC/000267.

References

1. Goñi-Menoyo, J.M., González-Cristóbal, J.C., Villena-Román, J.: MIRACLE at Ad-Hoc CLEF 2005: Merging and Combining without Using a Single Approach. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022. Springer, Heidelberg (2006)
2. Lana-Serrano, S., Goñi-Menoyo, J.M., González-Cristóbal, J.C.: MIRACLE at GeoCLEF 2005: First Experiments in Geographical IR. In: Peters, C., Gey, F.C., Gonzalo, J., Müller, H., Jones, G.J.F., Kluck, M., Magnini, B., de Rijke, M., Giampiccolo, D. (eds.) CLEF 2005. LNCS, vol. 4022, pp. 920–923. Springer, Heidelberg (2006)
3. Goñi-Menoyo, J.M., González-Cristóbal, J.C., Lana-Serrano, S., Martínez-González, A.: MIRACLE's Ad-Hoc and Geographic IR approaches for CLEF 2006. In: Peters, C., et al. (eds.) CLEF 2005. LNCS, vol. 4022, Springer, Heidelberg (2006)
4. Lana-Serrano, S., Villena-Román, J., Goñi-Menoyo, J.M.: MIRACLE at GeoCLEF Query Parsing 2007: Extraction and Classification of Geographical Information. In: Nardi, A., Peters, C. (eds.) Working Notes of the Cross Language Evaluation Forum (CLEF) 2007 Workshop, Budapest, Hungary (2007)
5. Zhisheng, L., Chong, W., Xing, X., Wei-Ying, M.: Query Parsing Task for GeoCLEF2007 Report. In: Nardi, A., Peters, C. (eds.) Working Notes of the Cross Language Evaluation Forum (CLEF) 2007 Workshop, Budapest, Hungary (2007)
6. Geonames geographical database, `http://www.geonames.org`
7. U.S. National Geospatial Intelligence Agency, `http://www.nga.mil`
8. U.S. Geological Survey, `http://www.usgs.gov`
9. Global 30 Arc-Second Elevation Data Set, `http://eros.usgs.gov/products/elevation/gtopo30.html`
10. Charniak, E.: A Maximum-Entropy-Inspired Parser. In: Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, USA (2000)
11. University of Neuchatel. Page of resources for CLEF, `http://www.unine.ch/info/clef`